Problems in Large-Scale Exploitation:

Making Very-Large into a Good Thing

DARPA-NGA Industry Workshop

Dr. Michael Pagels

7 September 2005

michael.pagels@darpa.mil

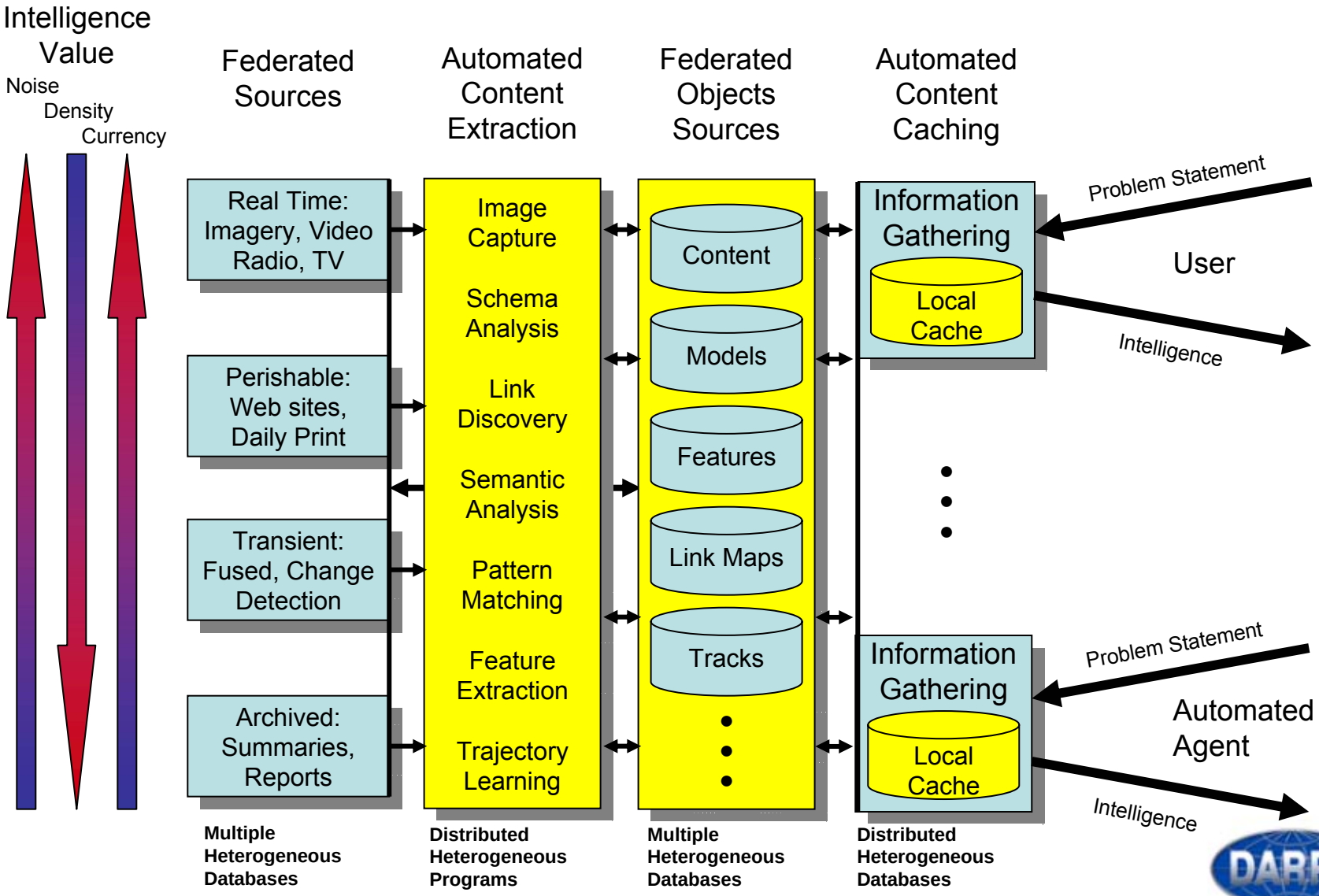
What is Large-Scale Exploitation?

- A capability for storing, retrieving and analyzing enormous quantities of data
- Use all of the new stuff we're about to collect, and 50+ years of old stuff
- Unify handling of radar and hyperspectral imagery, video and motion, and other sensor data
- Access derived information, from imagery, text, analyst annotations and others
- Retrieve items of interest very rapidly

Solution requires near fully automated processing

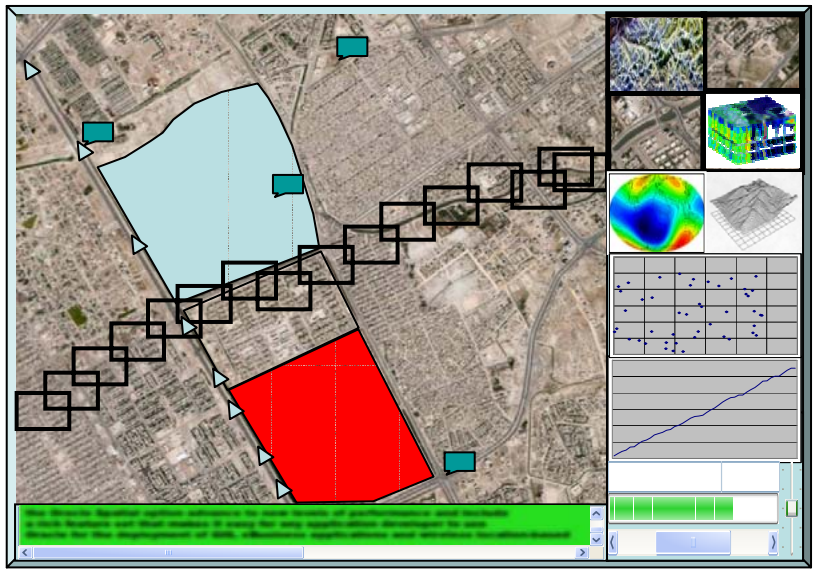
Large-Scale Exploitation:

We know the framework, we need revolutionary technologies



From This

To This



Research Theme Areas

A sampling of technology areas which need innovations to meet global scale problems

Important Problems

1. Automated multidimensional association

How is the raw data and extracted information, indexed and associated to support efficient exploitation?

- Unified registration and indexing across space, time, and relations

The extension and generalization of R*-trees to incorporate temporal content while not treating time as just another spatial dimension

- Unsupervised discovery of relationships in space and time
- Refinement and projection of associations from very few human interactions
- Derivation and maintenance of precision, accuracy and provenance metrics

Important Problems

2. Multiple model query language & interface

How is the indexed and associated information searched to glean useful pieces of intelligence?

- Google for Intelligence means an Exploitation Query Language; all-source unified query specifications of space, time and relations
- What is the analysts' access (HCI) to EQL
- New query optimization techniques

Utilize "just-in-time-reindexing" to generate a custom index of a portion of the database, specifically organized for an individual analysts mission.

- Utilize query "failures" to inform collection management of unfulfilled data requirements

Important Problems

3. Workflow-aware distribution & caching

How does the most useful subset of data get within reach of the user for efficient exploitation?

- Predict a user's data needs for the next 24h
- Stage a superset of data for rapid access
- Maintain coherence between all of the sets

Generalize and extend distributed database coherence protocols to be resilient to Byzantine error conditions with unpredictable comms bandwidth.

- Scalable across a diverse range of users

Important Problems

4. Automated model & view transformation

How do we automatically express information in a manner consist with the analyst's mission and access?

- Multiple sources, but a single portal
- Recast the user's view of the data based on the specifics of their mission
- Drive knowledge discovery from the specifics of the mission

Utilize machine learning to determine the relationships between mission objectives and queries, and use that knowledge to pre-stage queries.

- Maintain and apply multi-level security policies

Important Problems

5. Automated schema discovery + maintenance

How do we automatically discover and utilize new information sources and maintain access to old ones?

- New source discovery for structured and unstructured data
- Automated concept discovery and association

Utilize unsupervised probabilistic reasoning to construct a hierarchy of information clusters across multiple dimension of the data set.

- Automated schema alignment and maintenance
- Automated index “column” discovery

Important Problems: Research Theme Areas

1. Automated multidimensional association
2. Multiple model query language & interface
3. Workflow-aware distribution & caching
4. Automated model & view transformation
5. Automated schema discovery + maintenance

Scalable, fully automated, eliminate manual database support

“So, all that’s left is everything.”

Jesse James, *Monster Garage*

Approved for Public Release, Distribution Unlimited